

Automatsko indeksiranje i tf-idf mera

Zadaci

7.1.1 Za date dokumente, izdvojiti ključne reči koje identifikuju sadržaj dokumenta (samo imenice), posmatrajući:

- (a) apsolutan broj pojavljivanja (n_{ij} , nenormalizovane vrednosti) termina u dokumentu; Posmatrati samo termine sa apsolutnom frekvencijom pojavljivanja većom ili jednakom od 2. Prokomentarisati rezultate upita "apple", "apple AND gadget", "apple AND salad".
- (b) broj pojavljivanja termina u dokumentu u odnosu na ukupan broj značajnih reči u dokumentu (term frequency, tf_{ij} , normalizovane vrednosti). Posmatrati samo termine sa normalizovanom frekvencijom pojavljivanja većom ili jednakom od 0.25. Prokomentarisati rezultate upita "apple", "apple AND gadget", "apple AND piece", "apple AND fruit".
- (c) indeks značaja (inverse document frequency, $idf_j = \log_{10} \frac{N}{df_j}$, N - broj dokumenata u kolekciji)
- (d) združeni indeks (kombinovana mera, $tf - idf$, $w_{ij} = tf_{ij} * \log_{10} \frac{N}{df_j}$). Za date upite proveriti u kojoj meri dokumenti zadovoljavaju te upite: "apple AND gadget", "fruit AND salad", "apple AND iphone AND gadget".

D_1 : Apple has launched new gadget. Apple named it "apple".

D_2 : In order to prepare fruit salad, cut apples in pieces. Mix apples and banana pieces.

7.1.2 U Reuters kolekciji od $N = 806791$ novinskih članaka, date su frekvencije pojavljivanja sledećih reči (ukupan broj pojavljivanja reči u svim dokumentima):

termin	df_t
car	18 165
auto	6 723
insurance	19 241
best	25 235

Dat je i apsolutan broj pojavljivanja reči unutar dokumenata D_1 , D_2 i D_3 iz kolekcije:

termin	D_1	D_2	D_3
car	27	4	24
auto	3	33	0
insurance	0	33	29
best	14	0	17

Broj reči po dokumentima: $|D_1| = 4237$, $|D_2| = 11298$, $|D_3| = 2011$

- (a) Izračunati indeks značaja ovih termina za indeksiranje (idf). Koji je najznačajniji termin za indeksiranje prema ovoj meri?
- (b) Izračunati združeni indeks ($tf - idf$) za ova tri dokumenta iz kolekcije. Koja reči su najznačajnije za indeksiranje svakog od tih dokumenata?

7.1.3 U kolekciji od $N = 234458$ dokumenata, date su frekvencije pojavljivanja sledećih reči (ukupan broj pojavljivanja reči u svim dokumentima):

termin	df_t
obrazovanje	40 186
prosveta	22 182
fakultet	14 684
katedra	36 240

Dat je i apsolutan broj pojavljivanja reči unutar dokumenata D_1 i D_2 iz kolekcije:

termin	D_1	D_2
obrazovanje	26	10
prosveta	34	5
fakultet	53	14
katedra	41	7

Broj reči po dokumentima: $|D_1| = 2345$, $|D_2| = 864$

- (a) Izračunati indeks značaja ovih termina za indeksiranje (idf). Koji je najznačajniji termin za indeksiranje prema ovoj meri?
- (b) Izračunati združeni indeks ($tf - idf$) za ova tri dokumenta iz kolekcije. Koja reči su najznačajnije za indeksiranje svakog od tih dokumenata?

7.1.4 U kolekciji od $N = 123456$ dokumenata, date su frekvencije pojavljivanja sledećih reči (ukupan broj pojavljivanja reči u svim dokumentima):

termin	df_t
računar	42 000
pc	18 360
mac	26 780
monitor	10 250

Dat je i apsolutan broj pojavljivanja reči unutar dokumenata D_1 i D_2 iz kolekcije:

termin	D_1	D_2
računar	64	22
pc	20	40
mac	5	32
monitor	8	21

Broj reči po dokumentima: $|D_1| = 630, |D_2| = 840$

- (a) Izračunati indeks značaja ovih termina za indeksiranje (idf). Koji je najznačajniji termin za indeksiranje prema ovoj meri?
- (b) Izračunati združeni indeks ($tf - idf$) za ova tri dokumenta iz kolekcije. Koja reči su najznačajnije za indeksiranje svakog od tih dokumenata?